Emergent data analysis in phonetic sciences: Towards pluralism and reproducibility in phonetic sciences

Timo B. Roettger¹, Bodo Winter², Harald Baayen³

¹ Northwestern University, Department of Linguistics
² University of Birmingham, Dept. of English Language and Applied Linguistics
³ University of Tübingen, Department of Linguistics

Corresponding author
Timo B. Roettger
Northwestern University, Department of Linguistics
841 W Windsor Avenue, 60640 Chicago, IL
timo.b.roettger@gmail.com

Abstract
This special issue introduces a series of papers focused on not only discussing and making available new methods to the phonetic and linguistic community but also reflecting upon existing data analysis practices. In our introduction, we highlight how the contributions to this special issue fit within the landscape of statistical approaches, and how they are enmeshed with issues regarding the difference between exploratory and confirmatory analyses, issues regarding different approaches to statistical inference, as well as issues with the analysis of massively multidimensional multivariate speech data. Moreover, we provide a call for considering the importance of open and reproducible research practices, such as the publication of data and code that was used for a study’s data analysis. Rather than being dogmatic about particular statistical methods, the pluralism of analysis approach should excite debate and discussion, to which this special issue is an invitation. In addition, the co-existence of multiple ways of analyzing the same data (each with its own advantages and disadvantages and different analysis goals) makes it all the more important for researchers studying spoken language to make their research process open and more accessible to other researchers.

Keywords: data analysis, statistics, reproducible research, open science, significance testing, Bayesian modeling
1. Introduction

The landscape of data analysis in linguistics and other fields is constantly changing. Advances in computational power have made new analytical approaches possible, and the use of open access software such as R increases the speed with which new statistical methods are shared within our field and across disciplines. As accessibility to these methods increases, more and more people within linguistics employ increasingly complex analytical techniques to their data. Parallel to the ever-growing toolkit of statistical methods, there are shifts in methodological traditions and statistical philosophies, with an array of differing views about how data should be analyzed, how it should be reported, and how it should be shared. In sum, the field of data analysis is in flux. Amidst the backdrop of changing practices, it is important to critically assess past practices, to reflect upon present practices, and to look out for what new developments will affect our future practices.

We approach data analysis with George Box’s famous quote in mind, “all models are wrong, but some are useful” (Box, 1979: 2). This often-repeated quote embodies a fundamental truth about data analysis, which is that we perform analyses to gain a better understanding of our world and the phenomena we investigate, i.e., statistical models are supposed to be “useful” to the scientific community. On the other hand, all models are also “wrong” to some extent, with each model providing only a snapshot of the underlying complexity of the data to be modeled. Models can be “useful” in different ways and to differing degrees, and models can be more or less “wrong” as well. There is no single model that is the best model and that is equally useful across theories and phenomena. This very fact necessarily creates a plurality of analytical approaches, within and across disciplines. Even expert statisticians reach different conclusions when given the same dataset (Silberzahn et al., 2017). Rather than trying to provide gold standards and recipes, we endorse the plurality of approaches and highlight that pluralism calls for comparison, reflection, and a critical discourse about methods. We should not try to elevate any one method to the status of a “best” method or a canonical way to analyze particular datasets; instead we should discuss the advantages and disadvantages of particular approaches.

In line with the idea of plurality, data analysis varies along important dimensions. We would like to highlight a few of these dimensions to not only introduce the contributions to our special issue, but also to review what we conceive as important topics for data analysis in phonetic sciences. These relevant dimensions are the divide between exploratory and confirmatory analysis (§1.1), the divide between null hypothesis significance testing (NHST) and Bayesian approaches to inference (§1.2), and ways of tackling the multidimensionality of phonetic data (§1.3).

Beyond reflecting on past and future methods it is also important to think about how data analyses are communicated and shared with the community. Section 2 will discuss the relevance of reproducibility and discusses the benefits of an open and transparent phonetic community, exemplified by the contributions in this special issue.

1.1 Exploratory vs. confirmatory data analysis

It is important to recognize that data analysis includes two stages, which are more or less conceptually distinct, although they may overlap to considerable degrees in practice. In an exploratory stage, a researcher observes patterns and relationships leading to the generation of new hypotheses as to how these observations can be explained. We can label this stage “hypothesis-generating”. Many breakthroughs in science originate in the serendipity of researchers observing an unexpected anomaly while scrutinizing their data. In a confirmatory stage, novel hypotheses as well as hypotheses extending or challenging established theories are then pitted against new data, obtain in, for example, controlled experimental studies. We may label this stage “hypothesis-testing”. Putting our hypotheses under targeted scrutiny via confirmatory tests helps us to build up sufficient evidence for actually revising established theories. The revised theories can then be further informed by
additional exploration of the available data, leading to an iterative process that alternates between exploration and confirmation. Exploratory and confirmatory phases of analysis should be considered complementary, and they are both necessary components of scientific progress.

Unfortunately, when it comes to publishing work within phonetics and other areas of linguistics, exploration and confirmation are not weighted equally. Confirmatory analyses have a superior status, determining the way the way funding agencies demand that proposals should look like, and shaping how we frame our papers. The prestige of confirmatory statistics is so high that occasionally the review process can force authors to recast the reportage of exploratory analyses in the format of the reportage of confirmatory analyses (see, e.g., Pham & Baayen, 2015, footnote 1).

The emphasis on confirmation over exploration has large-scale consequences for published research. It is important to realize that in a true confirmatory setting, researchers have only one shot. Harrell (2015), a monograph - targeting medical statisticians - on regression strategies for confirmatory data analysis, allows for initial graphical exploration of the data, after which a single, theoretically motivated, model is fitted to the data that focuses on an a-priori established clearly defined hypothesis. Subsequently, model criticism is carried out to clarify whether the resulting model is actually appropriate for the data. In true confirmatory analysis, there is no place for repeated modeling during data collection, no place for adding or removing interactions, and no place for including or removing control variables. As soon as a second model is fitted to a given data set, the analysis is no longer confirmatory, but exploratory (see Baayen, Vasishth, Kliegl, & Bates, 2017, for further discussion).

Unfortunately, the results of what has actually been an exploratory analysis are often presented as if they were the results of a confirmatory analysis. The prevalent expectation that the main results of a study should be predicted based on a priori grounds has led to harmful practices for scientific progress (John, Loewenstein & Prelec, 2012), such as HARKing, “Hypothesizing After Results are Known” (e.g., Kerr, 1998). Additionally, researcher degrees of freedom in conducting studies and analyzing data can be exploited to hunt for significant p-values, also known as p-hacking (see also Simmons, Nelson, & Simonsohn, 2011), which can ultimately be reported as if confirming a planned analysis.

The point is not that nobody should perform exploratory analyses—quite to the contrary. The point is that each analysis needs to be clear about where it stands, and to what degree it is confirmatory or exploratory. Many studies in the phonetic sciences are exploratory in nature and should be treated as such. The complexity of speech naturally means that we do not always have specific directed hypotheses for all aspects of the data, and there are many interesting patterns to be discovered, and later confirmed on separate datasets. It is often the exploratory part of the analysis that we can learn the most from, especially with highly multidimensional data. For series of multivariate exploratory studies addressing complex linguistic data, consistent results for multiple predictors across independent data sets provides cumulative evidence for constellations of factors that goes beyond analyses targeting single specific hypotheses. Importantly, researchers carrying out exploratory data analysis can to some extent protect themselves and their colleagues against red herrings by setting much more stringent alpha-levels when evaluating whether there are signals in the noise (e.g. Benjamin et al., 2018). In exploratory analysis, it is the researcher’s duty to launch adversarial attacks on potential effects, and only report effects as of potential interest when attempts to bring them down fail consistently.

If a strict null hypothesis significance testing approach is followed, confirmation cannot happen on the same dataset that was used as the basis for exploration. To the extent that confirmatory and exploratory analyses may blend into each other in actual practice, the researcher needs to be aware of this and report results accordingly.

1.2 Inferential frameworks: Frequentist and Bayesian inference
An important aspect of data analysis is making generalizable statements about observations. Inferential statistics is the process of using samples to make "inferences" for parameters of a population of interest. For example, a study may contain a subset of speakers from a linguistic community, and the sample is used to make inferences about all speakers of the language. Or a study may contain a subset of words from the language, and the sample is used to make inferences about all words of the language. In statistics, there are two different approaches to making this inference, known as frequentist and Bayesian statistics, and there has been considerable debate about the relative merits of each approach. Each approach has different analysis goals and makes different assumptions. Our special issue includes several papers that discuss aspects of different inferential frameworks as well as papers that make use of techniques and methods developed within these frameworks.

Classical methods for statistical inference (analysis of variance, discriminant analysis) are grounded in the work of Sir Ronald Fisher (1925). These methods, which are widely used in phonetics and many other fields of inquiry, are known as frequentist, as they are grounded in a particular understanding of the concept of probability, namely, the idea that the probability of an event is given by the limit of its relative frequency across a large number of trials. Fisher’s method was later combined with Neyman and Pearson’s approach to hypothesis testing (1928) to create what is now known as null hypothesis significance testing (NHST) (Lindquist, 1940; Gigerenzer, Kraus, & Vitouch, 2004). These approaches to inference allowed researchers to evaluate whether for their data there was signal in the noise or not, resulting in an extremely useful tool at a time that computers did not exist, and that has been used ever since across the sciences. Traditional NHST works by assuming a null hypothesis (there is no signal, only randomness) and gathering evidence against that initial assumption. The $p$-value measures the incompatibility of the data with the null hypothesis, and it is often used as a hard cut-off, where an effect is accepted as “significant” if its associated $p$-value falls below a preset threshold probability. The adherence to the $p < 0.05$ threshold probably grew out of Fisher's very practical advice to not take seriously effects for which $p > 0.05$ (which was in part motivated by the 2.5% cutoff-points of the normal distribution being close to -2 and 2 in probability tables). NHST (as per Neyman and Pearson) provides a simple and specific decision procedure (using a particular threshold, such as $p < 0.05$) which will assure low error rates in the long run, across a series of repeated experiments.

The practice of NHST has been much criticized by researchers, among other things for relying on an arbitrarily defined hard-cut threshold for “significance” (rather than taking the continuous strength of evidence into account), and for overly emphasizing point estimates (such as means) over interval estimates (such as confidence intervals or credibility intervals) (e.g., Cumming, 2012, 2014). Moreover, whenever NHST is applied to a new data set, the framework is deliberately naïve about how much we know about the object of study. We note here that classical ‘frequentist’ inference is not necessarily or intrinsically focused on hard cut-offs, and that a $p$-value and confidence intervals can be used profitably without having such a cut-off in mind. What appears to have made NHST so attractive to many researchers is that it provides a clear criterion for getting results published, and hence has become a tool that can be, has been, and is being misused to advance one’s career.

Frequentist inference, as introduced by Fisher, differs in many ways from what is now known as Bayesian inference. The field of statistics has a long history of a deep divide between classical ‘frequentist’ statistics and Bayesian statistics, each camp having its own philosophical foundations and methodological goals, and each camp completely rejecting the validity of the philosophical underpinnings of the other camp. Nowadays, many statisticians feel the debate has outlived its usefulness, and the emphasis is shifting towards developing statistical methods that help researchers understand and evaluate their data, including the uncertainty about model estimates and the importance of the accumulation of knowledge.

The term Bayesian covers two classes of models, approximate Bayesian models and standard Bayesian models. Standard Bayesian models give the analyst the freedom to posit prior distributions for all model parameters of interest. These models make use of sampling techniques to estimate the posterior distributions of these model parameters, as posterior
distributions often cannot be calculated or approximated analytically. The interest of the analyst will typically focus on how the posterior distribution is positioned with respect to zero, and whether a given observed parameter value is surprisingly extreme given its posterior distribution. The paper by Vasishth et al. (this volume) provides an introduction to standard Bayesian analysis.¹

Approximate Bayesian models are models that include priors, but estimate the priors from the data, using asymptotic approximations to save computational time. These models are known as Maximum a Posteriori (MAP) models and also as empirical Bayes models. The generalized additive model (GAM) as implemented in the mgcv package for R, introduced in the paper by Wieling (this volume) is a MAP model that follows the empirical Bayes approach in which the prior parameters are determined from the data itself. Quantile GAMs (Fasiolo et al. 2016), generalized additive models that predict the median, deciles, and other quantiles of interest, rather than just the mean, are empirical Bayes combined with asymptotic approximations in combination with a loss rather than a prior.

In a standard Bayesian approach, the prior should be fixed before the data is observed, something that is impossible in GAM modeling. However, if one sees the prior as part of the model (which is probably a sensible point of view in GAM modeling), one can see the empirical Bayesian posterior as an approximation to the standard Bayesian posterior, where some parameters have been fixed to the maximum of the marginal likelihood. Thus, the confidence intervals provided by GAMs are Bayesian ‘credible’ intervals, even though they are not advertised as such. For those who want to go ‘fully’ Bayesian, the jagam package for R (Wood, 2016) allows the analyst to set up fully Bayesian models from GAM models specified with the mgcv package, which makes practical sense only when the goal is to extend the GAM model with further parameters that go beyond what the optimized algorithms of mgcv can offer.

The philosophical underpinnings of standard Bayesian inference are very different from those of frequentist inference and the particular codification thereof known as NHST (for a comparison of NHST and Bayesian inference, see Dienes, 2008). Standard Bayesian modeling provides a general, very flexible approach to statistical modeling, allowing researchers to test models that are analytically intractable. Furthermore, standard Bayesian modeling makes it possible for the analyst to include knowledge gleaned from previous research when modeling new data – the Bayesian paradigm has, intrinsically, a much more cumulative perspective on the gathering of scientific evidence. Classical inference can be, and all too often is, focused far too much on the significance of individual tests, forgetting that solid evidence emerges only out of series of replication studies. Consider, by way of example, a series of 10 experiments, each of which fail to provide a significant effect for a given factorial contrast. Assume, furthermore, that for 9 out of 10 tests, the mean is greater than zero. Rigid application of the NHST procedure would lead one to conclude there is no effect, but a simple sign test will clarify immediately that in all likelihood the mean is positive. The Bayesian paradigm offers much more sophisticated tools for integrating knowledge across multiple studies. Yet these tools come with more researcher degrees of freedom and can place the analyst for new dilemmas. Consider, for instance, whether information about mean and variance of an effect obtained in a previous study should be incorporated into the model one is about to fit to a given dataset. In a strict confirmatory setting, this decision has to be made before carrying out the analysis. What one cannot do in a confirmatory analysis is fit the model to one’s own data, observe there is good reason to suppose an effect is present, then run a model with priors taken from the preceding study, observe the evidence for the previously observed effect has evaporated, and then discard the evidence from the preceding study on the grounds that it is potentially flawed or incompatible.

The Bayesian paradigm is also in general adverse to providing rigid cut-off points for “statistical significance”, and calls attention to the uncertainty associated with parameter

¹ We note here that a fully Bayesian framework can be based on a loss function, rather than a prior, see Bissiri et al. (2016) for detailed discussion.
estimates. As a consequence, it can leave the researcher with no hard and fast criteria for assessing whether, for a given experiment testing a novel hypothesis, there is reason for surprise. Here, at the end of the day, the Bayesian statistician will have to provide researchers with an answer to the question whether there is good reason to suppose an effect is present. Just as Sir Ronald Fisher, when asked what should count as significant, came up with the $p < 0.05$ rule of thumb, the Bayesian statistician has to provide the researcher with some guidelines as to whether an effect is ‘significant’, in the sense that the effect is clearly worth exploring in further research, or offers useful confirming evidence.

One striking difference between frequentist and approximate Bayesian models on the one hand, and standard Bayesian models on the other, is that the latter confront the analyst with many more options and choices than the former. In modern parlance, the standard Bayesian analyst comes with many more researcher degrees of freedom. This greater freedom is wonderful on the one hand, as it makes it possible to develop more sophisticated models. But on the other hand, this greater freedom comes at the cost of follow-up research making different choices and setting up their models in slightly different ways. Sir Ronald Fisher is known for having been critical of Bayesian statistics – the term Bayesian was coined by him (see Fienberg, 2005) - precisely because conclusions can change depending on the choice of prior. The tradition of NHST grew out of the need of having a clear-cut procedure for evaluating the results of statistical tests. That the need for clear ‘gold standards’ is still felt today is exemplified by the paper by Barr et al. (2013) on how to fit mixed models. With the wide spectrum of analytical techniques currently available, which will increasingly also include methods from machine learning, it is not possible nor desirable to enforce rules by means of which significance can be assessed mechanically. A spirit of plurality is needed that creates space for realizing that there are problems and applications that might be handled more easily by either Bayesian or frequentist approaches, and that the analyst is far better off having both tools in his toolbox.

There are two papers in our special issue that illustrate the strengths and opportunities offered by standard Bayesian approaches. Vasishth et al. (this issue) give an extended overview about the logic and benefits of standard Bayesian analyses and walk the reader through a concrete standard Bayesian analysis of an acoustic study, investigating whether and how voice onset time measurements discriminate different stop series across six languages. Nicenboim, Roettger, and Vasishth (this issue) investigate the phenomenon of incomplete neutralization of German final devoicing, a phenomenon that has sparked extended debates regarding the available evidence (Port & O’Dell, 1985; Fourakis & Iverson, 1984; Roettger et al., 2014).

One reason for the skepticism about the phenomenon of incomplete neutralization is that it is at first sight unclear to what extent experiments reported in the literature generalize. Many studies investigating incomplete neutralization based their conclusion on very small samples of speakers. The smaller the sample, the less likely do our results generalize beyond the observations made. This issue falls under the header of statistical power, the probability that the statistical test will reject a false null hypothesis in a NHST framework. Kirby and Sonderegger (this issue) show what sample size is required to accurately detect any small effect. They utilize numerical simulations to generate hypothetical datasets; this allows characterizing how large a sample has to be in order to detect a real effect. In line with these findings, Nicenboim et al. (this issue) perform a meta-analysis on the available literature and show that, cumulatively, there is substantial evidence for the claim that final devoicing in German is incomplete, although many of the individual studies alone do not allow this conclusion because of their limited sample sizes. Together, these papers highlight the issue of statistical power (for a discussion and software, see also Westfall et al., 2014), as well as the issue of evidence synthesis across studies to circumvent basing interpretations on low powered individual studies.

We want to emphasize that in the spirit of plurality, we do not intend the inclusion of these papers to suggest that analyzing data with frequentist or empirical Bayes methods are wrong. Given the prevalence of the decision procedure of NHST inside phonetics, we think that it is most prudent at this stage to be aware of the uncertainties that come with different
statistical methods, and the opportunities offered by standard Bayesian inference to address these uncertainties. Learning about standard Bayesian methods may also help clarify misunderstandings about NHST (see Nicenboim et al., this issue).

In addition to two standard Bayesian papers, the present volume also includes two papers that introduce and discuss an empirical Bayesian modeling tool, the generalized additive (mixed) model. The Generalized Additive Model (GAM) is an extension of the classical generalized linear model (GLM), which enjoys wide use within phonetics (e.g., multiple regression, logistic regression, linear mixed effects models). GAMs extend GLMs with methods for modeling smooth nonlinear functional relations between a response and one or more predictors (Wood, 2006; Winter & Wieling, 2016). GAMs can also incorporate random-effect factors. In addition, GAMs offer tools for addressing autocorrelations in the residual error, which are often present in time-series data - when observations are ordered in time, current observations may depend on previous observations.

Wieling (this issue) introduces GAMs and offers a step-by-step tutorial based on an analysis of articulatory data (for other introductions, see, e.g., Winter & Wieling, 2016, and Baayen et al., 2017). In the other paper that is focused on GAMs, Sóskuthy (this issue), a series of simulations is reported that explore the impact of a range of modeling options that researchers have when using GAMs. As with any new tool, it is not always clear what the best approach to using this tool is from the get-go. This was the case with linear mixed effects models, which incited a prominent debate about what the best random effects structure for the analysis of experimental designs is (see Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). The flexibility inherent to statistical modeling is amplified in the case of GAMs, which provide many more options to their users. For a set of simulated trajectories, Sóskuthy looks at the impact of these degrees of freedom on how often we erroneously reject / accept the null hypothesis under certain assumptions.

1.3 Dealing with the multidimensionality of speech communication

Our choice of data analysis also varies tremendously as a function of the way phenomena are observed and measured. Dependent on how observations are operationalized, certain analytical tools may or may not apply. Speech is inherently multidimensional and varies across time, as is the case with pitch curves, formant trajectories or articulatory gestures. This time-series data can be analyzed as a sequence of static landmarks (“magic moments”, Vatikiotis-Bateson, Barbosa, & Best, 2014) or as continuous trajectories, depending on how relevant the dynamic nature of speech behavior is for any given theory (Mücke, Grice, & Cho, 2014). The papers by Wieling (this issue) and Sóskuthy (this issue) tackle this aspect of data analysis and discuss the value of GAMs to model time-series data.

Plummer and Reidy (this issue) discuss another issue related to the multidimensionality of phonetic data analysis. They discuss a method for computing low-dimensional representations of speech. The method they introduce centers on the use of Laplacian Eigenmaps to build structures over data points from which low-dimensional representations of speech are learned. This technique enables us to reduce the multi-dimensional acoustic signal to lower dimensionality, which, as they argue, are a better proxy of cognitive and social speech categories.

Another aspect of multidimensionality is tackled by Tomaschek, Hendrix, and Baayen (this issue). They deal with a common problem in regression analyses (and by extension mixed models, GAMs etc.), namely, the issue of collinearity. When predictor variables in a model are highly correlated, estimating parameters may become severely unstable and researchers can easily draw the wrong conclusions based on their data. Collinearity is an important problem that is often overlooked. As stated by Zuur, Leno, and Elphick (2010: 9): “If collinearity is ignored, one is likely to end up with a confusing statistical analysis in which nothing is significant, but where dropping one covariate can make the others significant, or even change the sign of estimated parameters.” Tomaschek et al. (this issue) introduce the
reader to new methods that allow coping with collinearity. Their discussion is relevant for data analysis situations in which there are lots of highly correlated variables to consider.

Another aspect of multidimensionality is the non-verbal context in which speech occurs. Speech communication does not happen in a void, but is accompanied by changes in body posture, head position, gaze, facial expressions, and manual gestures (McNeill, 1992; Goldin-Meadow, 2003; Kendon, 2004). Danner et al. (this issue) invite the reader to rethink how to characterize multimodal speech by applying dynamic approaches already used in speech research to multimodal communication. They discuss both the problem of automatically identifying visual gestures in video images, as well as the problem of correlating a gestural data stream with an acoustic data stream.

In the spirit of “endorsing plurality”, the papers discussed so far are either focused on the merits and pitfalls of different statistical philosophies (such as NHST versus Bayesian inference), or they discuss various new methods that are useful for different phonetic applications. Another strand that runs across the entire special issue is the issue of reproducibility. Reflecting on methods does not end with choosing a particular method, but it also includes thinking about how data analyses are communicated and shared with the community.

2. Towards reproducible phonetic sciences

Many things increase or decrease the credibility of a scientific finding (Thagard, 1978), but chiefly, the evidence supporting a theory is a major factor. To assess the strength of evidence for a theory, one needs to take into account how the data was collected, and how it was analyzed. Evaluating the strength of evidence becomes very difficult if part of the research process is not transparent. Reproducible research involves the capacity of other researchers (who have not conducted the original study) to repeat the analysis that is presented in a published study (see Peng, 2011; Munafò et al., 2017). Reproducibility minimally necessitates that the raw data and the analysis code is made available to the community if this is possible.

For our field, reproducible research has numerous advantages. First, as mentioned above, even expert data analysts will perform different analyses based on the same dataset (Silberzahn et al., 2017). Naturally, different analysis choices yield different conclusions. Each analysis is characterized by a “garden of forking paths” (Gelman & Loken, 2014) or what Simmons et al. (2011) call “researcher degrees of freedom”. Some relevant researcher degrees of freedom for phonetic studies include what data we exclude from our data set prior to statistical analysis; whether clusters of observations (speakers, listeners, words, etc.) are aggregated over or not; what predictors out of numerous possible predictors are included in an analysis, etc. etc. (for a discussion of researcher degrees of freedom in phonetics, see Roettger, 2018). McElreath (2016) emphasizes that statistical modeling is subjective, in the sense that it incorporates the researcher’s beliefs and assumptions about a study system. Because of this, the only way to allow evaluation of the process of statistical modeling by outsiders is to make it open². This transparency then allows other researchers to draw their own conclusions based on the same dataset, reanalyze other aspects of them etc.

Publishing the raw data and code also facilitates knowledge transfer: Other researchers can learn from the ways a particular dataset was analyzed, and how the analysis was implemented in actual software code. It is with this spirit of sharing knowledge

---

² At present, many of the descriptions of statistical methods found in phonetics papers do not allow reproducing the performed analysis; in some cases, it is not even clear what general analysis was conducted (e.g., p-values may be listed without a detailed description of the associated statistical models these values are based on). For example, Winter (2011) tried to assess how frequent the independence assumption is violated in speech production data and found that many publications in phonetics do not provide enough information to allow such an assessment. This is a dire situation to be in, as statistically minded readers are neither able to reproduce the analysis, nor are they able to evaluate the actual uncertainty surrounding the presented evidence.
and being transparent in mind, that all authors of this special issue make their code and data available on public repositories, allowing the readership of the special issue to readily implement the methods, as well as to actively participate in the discourse that surrounds the methods presented here. Reproducibility runs as a prominent thread through all of the papers in this special issue.

Many of our papers are written in a tutorial-like way, inviting the reader to reproduce and extend the offered analyses. For example, Politzer-Ahles and Piccinini (this issue) discuss ways to visualize the results of hierarchical models that allows to communicate the population-level estimates alongside the random variation associated with crossed random effects. Data visualization is an important aspect of communicating our findings and has been the subject of ongoing debates across scientific fields (e.g., Tufte, 1990; Kosslyn, 2006, Weissgerber et al. 2016). Politzer-Ahles and Piccinini’s paper not only serves as a reminder of the importance of data visualization in communicating data and statistical models, it also is a learning platform as their scripts allow ready application to new datasets.

The topic of reproducibility is also a prominent theme for Jadoul et al. (this issue). As argued by many proponents of reproducible research, all aspects of the research workflow interact with reproducibility, not just the “final” data analysis stage. For example, in acoustic analyses, there are many degrees of freedom as to what acoustic parameters to extract and how, such as which range settings to use for the measurements of a particular speaker’s fundamental frequency. We usually perform these analyses in available software such as Praat (Boersma & Weenink, 2018). However, data extraction in Praat is usually detached from subsequent statistical analyses. To streamline these processes, automated techniques can be used, for which Jadoul et al. (this issue) propose a new toolkit, Parselmouth, which integrates the extraction of Praat-based acoustic analysis into a Python-based workflow. For users of Python, this allows the combination of acoustic and statistical analyses within one script. For those who currently use Praat, Parselmouth may provide a useful alternative to streamline the process of acoustic analysis and make it more reproducible.

3. Conclusions
To conclude, we want to emphasize the spirit with which this special issue was conceived. As statistics is constantly evolving within and outside of linguistics and phonetics, there is a plurality of different analysis approaches. Many analytical philosophies alongside methodological tools and techniques co-exist alongside each other at any given point. In many ways, this is advantageous, as this creates the opportunity for discovery of new methods, many of which come from other fields, as well as the opportunity for honest discussion of the advantages and disadvantages of existing approaches. We are in no position, and nor is it our intention, to “police” any existing practices, or to provide recipes or guidelines that everybody should adhere to. Any strict rule will prove to be obsolete in the constantly changing landscape of statistical analysis anyway. Instead, we want to invite the community to reflect on existing practices (e.g., concerns regarding statistical power, null hypothesis significance testing, collinearity), as well as to look ahead to incorporate new analysis methods. Rather accepting any of these techniques as absolute, we have to continue the methodological debate as a community. Moreover, by becoming increasingly reproducible, we can ensure that this plurality of methods benefits our common scientific goal, to understand the physical, cognitive, and social aspects of human speech communication.
References


Kosslyn, S. M. (2006). *Graph design for the eye and mind*. OUP USA.


